

Developing a Data Enclave for Sensitive Microdata

Norman Bradburn
Senior Fellow
Education & Child Development Department

Randy Horton
Director of Development, Technology Services
Information Technology Department

Julia Lane
Senior Vice President
Economics, Labor, and Population Studies Department

Michael Tilkin
Senior Vice President and Chief Information Officer
Information Technology Department

1. Introduction

US federal agencies and departments disseminate statistical data to external researchers for a number of reasons. These include allowing external researchers to conduct analyses in areas of interest to the agencies, to identify data quality issues, or to find new and innovative research and educational uses for existing datasets without further burdening respondents or increasing agency costs.

To disseminate public use data, federal agencies can use a number of established commercial and academic archives, such as the Inter-University Consortium for Political and Social Research at the University of Michigan. However, there is a more limited range of options available to federal entities seeking to disseminate sensitive microdata that have not been fully de-identified for public use. The largest federal statistical agencies (e.g. Census Bureau and Bureau of Labor Statistics) have sufficient economies of scale to develop advanced in-house solutions that serve the needs of external researchers. Smaller agencies often lack the resources to archive, curate, and disseminate the datasets that they have collected.

This paper describes an approach to developing a statistical data enclave. The core features of such a data enclave are the curation, indexing and archiving of microdata; the provision of researcher access to sensitive microdata; statistical protection; researcher training; dissemination of information to the research community; the simultaneous support of a wide range of agency-specific data requirements; and the ongoing interaction with the federal consortium.

2. Background

The growth of data collection, particularly on humans and human organizations, which has been enabled by advances in cyberinfrastructure, presents a challenge to confidentiality protection, highlighted by a recent National Science Foundation workshop¹. Breaches of confidentiality that result from the actions of one researcher threaten the ability of scientists everywhere to collect and use data. Yet, preserving access to high quality scientific data is essential to the empirical replication that is at the core of good science. Indeed, as George Duncan indicated “In important cases social scientists [as well as scientists in the biological and environmental sciences] require the rich and sensitive data that makes confidentiality consequential. The technological factors that have led both to the explosive growth in the capability of providing such data, as well as in the capability of data snoopers to breach confidentiality, are in the domain of computer scientists.” (SBE/CISE workshop, March 15-16, 2005)

There is already an existing community that has focused on the importance of confidentiality. In particular, federal statistical agencies have devoted substantial resources to both statistical and technical ways to protect confidentiality², the Social and

¹ NSF SBE/CISE workshop, March 15-17, 2005, Airlie House, Virginia

² *Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies*, Pat Doyle, Julia Lane, Laura Zayatz and Jules Theeuwes, North Holland, 2001

Behavioral Research Working Group recently drafted a report entitled “Achieving Effective Human Subjects Protection and Rigorous Social and Behavioral Research” for the Human Subjects Research Subcommittee of the Committee on Science, National Science and Technology Council. Similarly, PITAC recently issued a report on cybersecurity that addressed some confidentiality issues and numerous studies have been undertaken by the National Academy of Sciences and the Committee on National Statistics.

Protecting databases against intruders also has a long history in computer science (a classic article is Dobkin, Jones and Lipton, 1979). Computer scientists themselves are interested in protecting the confidentiality of the data on which they do research (for example, the Abilene Observatory supports the collection and dissemination of network data, such as IP addresses).³ Cyberinfrastructure advances have certainly served to expand the set of access modalities, particularly with respect to remote access. The cybertrust initiative at NSF has created an entire research community that focuses on creating network computers that are more predictable and less vulnerable to attack and abuse, that is developed, configured, operated and evaluated by a well-trained workforce, and that educates the public in the secure and ethical operation of such computers. The Department of Defense has developed different levels of web-based access ranging from unclassified (nipr-net) to secret (sipr-net) to top-secret (jwics-net)⁴ using off the shelf technology. Similarly, the PORTIA project focuses on both the technical challenges of handling sensitive data and the policy and legal issues facing data subjects, data owners and data users.

Thus, as both Joan Feigenbaum and David Clark pointed out at the NSF workshop, the cyberinfrastructure advances that permitted the advances in data collection could also be harnessed to permit advances in data protection.

“Consequently my question about control of malicious behavior is not: can we build a solution, but rather: what is the right thing to build? To answer it, we must take into account a broad range of social and economic policy issues. All of us, each from our own disciplines, must come together and act as engineers to design this social artifact. In the cases where there is no tension among the stakeholders (e.g. where the problem is strictly technical), then the techies should just fix it” (David Clark, SBE/CISE workshop, March 15-16, 2005)

“We should strive to accompany sensitive data by easy-to-understand and easy-to-enforce policy metadata. Note that there has already been some good work on “privacy policies,” (e.g., the P3P work of the World Wide Web consortium – see <http://w3c.org/P3P>), but there has been little or no work on pushing these policies through all of the information technology that sensitive data encounter throughout their lifetimes and making sure that these policies are enforced.” (Joan Feigenbaum SBE/CISE workshop, March 15-16, 2005)

³ <http://abilene.internet2.edu/observatory/>

⁴ We are grateful to Carl Landwehr for making us aware of this.

How might cyberinfrastructure investments advance the protection of confidentiality? There are several key elements: addressing technical issues (e.g., building the right physical infrastructure); addressing social issues (developing protocols that people and organizations will use and implement security); developing trust (like Verisign); and continuing research agenda.

Building the appropriate physical infrastructure poses several clear scientific and engineering challenges, among them developing:

1. Usable interfaces to support a broad variety of personal privacy policies, and to express trust decisions;
2. Methods for storing, sharing, and aggregating data that are meaningful and yet respectful of integrity and privacy; and
3. Effective forensic tools for detecting and tracking malevolent activity, yet respectful of privacy.

Successfully addressing each of these challenges requires close collaboration between those expert in expressing the policies needed and those expert in understanding what policies are implementable.

3. The Approach

The core features of the data enclave should be the curation, indexing and archiving of microdata; the provision of researcher access to sensitive microdata; statistical protection; researcher training; dissemination of information to the research community; the simultaneous support of a wide range of agency-specific data requirements; and the ongoing interaction with the federal consortium.

Agencies should be able to work from a standard set of “checklists” to customize their offerings on both a per-agency and per-data source basis. These offerings could be customized in terms of areas such as data access protocols (e.g. onsite and Web-based), training options for external researchers and degrees of statistical protection used in the data.

3.1. Data Archiving, Indexing, and Curation

The data enclave should provide its partner agencies and departments with a set of offerings related to the archiving, indexing, and curation of data. Collectively, these offerings should ensure that these microdata are available for discovery and re-use over the long term by external researchers.

A first step in this process would be for the federal agency or department to deposit a master copy of a data set at the data enclave. To facilitate this process, the data enclave should prepare a document that clearly details both the guiding principles and standard options for archiving a data set. By placing a strong emphasis on understanding both the actual data and the agency’s requirements for the curation of the data, the data enclave

would be in a much better position to support external researchers on an ongoing basis, while minimizing the ongoing support required from the agency.

Using standard checklists that detail the available parameters for providing external researchers with access to the data set, the depositing agency could determine which of the data enclave's service offerings to utilize. The following are examples of business rules that could be determined through the checklists:

- Will the data enclave need prior approval from the agency before granting access to an external researcher who meets certain pre-defined requirements?
- Who at the agency will be the data enclave's representative responsible for approving all requests for data access?
- Will external researchers be able to create new data sets by combining data from multiple datasets housed in the enclave (including both data sets from a single agency and those from more than one agency)? What rules that will govern this process?
- Similarly, will external researchers be able to create new data set by combining enclave data with data from other sources and, if so, what are the rules will govern this process?
- In addition to the data enclave's standard data access authorization form, will the external researcher sign an agency-specific or data set-specific authorization form?
- Will external researchers be allowed to remotely access enclave data over the internet? If remote access is permitted, are there additional restrictions that will be placed on the data before this can occur?
- Will external researchers be permitted or required to complete an agency-specific or data set-specific training module before being allowed to access data?

All decisions related to the business rules applied to each data set would be completely under the agency's discretion. However, the data enclave should play an important value-added consultative role in this process. For example, the data enclave could advise the agency on the relevant best practices used by the data enclave's other customers, as well as by other statistical agencies and data enclaves both domestically and internationally.

In parallel with completing the checklist, the data enclave and depositing agency would need to work together to ensure that the data's structure and content (i.e. metadata) are clearly documented, using a standard, comprehensive set of metadata documentation that should be applied to all data in the enclave. For example, the data enclave and the depositing agency should fully document all variables (and combinations of variables) that are potential identifiers, and how they might be combined with other data sets to

reveal identities. This documentation would be used as a key reference in later de-identification efforts.

Once the data have been fully documented, the enclave would transform the data and metadata from their source formats and structures into a standard data enclave data format and structure -- the Data Documentation Initiative's (DDI) XML standard. By maintaining all of the data enclave's statistical data in a common format, the data enclave would streamline the delivery of this data to external researchers over the long term.

The data enclave could potentially also offer a rich set of online data discovery tools to both prospective and privileged researchers, providing them with the ability to interactively explore the metadata, and better understand what data is available to address their research questions. This functionality would help the federal entities "market" their available data sets to potential external researchers, as well as improve the sophistication and quality of the eventual data use. The enclave could leverage the common data format described above to populate these online data discovery tools directly from the data set's metadata.

An important function of the enclave would be to streamline the process of fulfilling external researcher requests. One way in which this could be done is by identifying situations in which multiple researchers request access to similar data products. For example, researchers may regularly request a custom data product that joins together the same two source data files. In these cases, it may be cost effective for the data enclave to invest some up-front effort in automating the creation of this custom data product.

The data enclave's long-term service offerings should go beyond providing access to statistical microdata and include the management of other types of sensitive research data, such as audio, video, image and unstructured text. This development would be aligned with data developments in social science research more generally, which is increasingly able to capture unstructured as well as structured data.

3.2. Provision of Research Access

The development of a robust set of data access tools that facilitate high-quality researcher interaction with the data, while at the same time ensuring that data confidentiality, would be critical. Two main data access modalities offer the most promise: an onsite access mode and a remote access mode that uses the Web to deliver the data. The tradeoffs between the modes involve balancing the convenience to an external researcher of working from his/her home office with the ability of the data enclave to more closely monitor the usage of the data.

With the remote access mode, the data enclave could provide external researchers with the ability to access anonymized data in a controlled manner over the internet. When a researcher needs to remotely access the data enclave's online resources, he/she could first initiate an encrypted connection with the data enclave using Citrix's web-based technology to access the applications and data provided by the data enclave. This

technology enables the data enclave to prevent an outsider from reading the data transmitted between the researcher's computer and NORC's network. Before the connection can be completed, the user must provide a pre-defined user id and password. In addition, using smart card technology, the user could also potentially be required to validate his/her identity in real time using an additional identification code.

The onsite data access mode would involve physically going to a secured site. At that site, preventive steps would need to be taken to restrict a researcher's ability to make illicit copies of the sensitive microdata and remove them from the premises. For example, the input/output devices on the site's computer would be locked down so that data files could not be transferred to an external device such as a USB flash drive. Similarly, the computer would not have access to any part of the network unrelated to the data enclave or to the public internet, preventing data from being illicitly transferred via the network. Any access to printers could be intermediated by data enclave staff. Finally, the data enclave would reserve the right to restrict and review all items entering and exiting the data analysis room. Should it become a requirement, the data enclave could also install a video camera that would allow the staff to monitor and record all activities in the data analysis room.

Through the Citrix technology, the data enclave would be able to deliver the same set of rich tools to both onsite and remote users. Every researcher would be set up with a private workspace. Since a major purpose of the data enclave is to facilitate productive, high-quality usage of microdata, this workspace would be designed to support the most useful elements of traditional, hands-on data analysis. For example, quantitative social science research is often a team-based collaborative activity that can involve multiple investigators supported by graduate students and research-funded staff. To the extent to which the customer agencies and departments allow multiple people on a team access to the data, team members can be set up with individual workspaces that are complimented by team workspaces. Each workspace would allow the user to save their result sets and related notes. Later development could include functionality that supports the ongoing collaborative annotation of data analysis and results, as well as the collaborative development of research deliverables such as journal articles. Finally, the enclave should expand future releases to support a set of data analysis tools built on the Online Analytic Processing (OLAP) technologies that are often used in data warehousing solutions.

Through the Citrix technology, researchers could be provided with a variety of powerful data analysis tools, such as statistical data processing packages (e.g. SAS, SPSS, or STATA). Researchers could also potentially utilize visual tools that support online data and metadata searching, browsing and analysis (e.g. NESSTAR). To enable high-quality use of the data, all data would be delivered in tandem with the related metadata, providing the researchers with the necessary context for data analysis.

The computing infrastructure should be built around a project-centric paradigm, where each project maintains a set of confidential data that can only be accessed on a "need to know" basis. This paradigm is a natural fit for the data enclave's computing needs. The enclave should create a partitioned "data enclave" zone residing on host computing

infrastructure. Only staff directly involved with supporting the data enclave would be able to access these resources. The computing security should also be validated through security audits conducted by qualified third party consulting organizations.

3.3. Statistical Protection

To access sensitive microdata, the external researchers would need to explicitly commit to protecting data confidentiality and subject themselves to the relevant federal confidentiality laws and their related penalties for violation. As a complement to this, the enclave would need to present agencies with a set of options for statistically protecting the confidentiality of the data before presenting the data to external researchers.

At a minimum, the data enclave should protect every data set by constructing a set of unique identifiers that can substitute for variables that are explicit personal/organizational identifiers, such as name, address, phone number, Social Security Number and Taxpayer Identification Number. Although there are a number of statistical protection techniques that perturb the data by means of noise addition, record swapping, rank swapping, blanking and imputation, the data enclave should work with each agency to determine the level of need for these techniques since they can be resource intensive and ultimately reduce the utility of sensitive microdata to external researchers.

Finally, the data enclave should limit researchers' access to the data they need for their specific research questions. To accomplish this, the data enclave can create custom analytic data files that contain a subset of the columns (and even rows) contained in the master data set.

3.4. Researcher Training

In keeping with the findings of the Portia project, it is important to develop the human as well as the physical infrastructure for the data enclave. Before being given access any data, external researchers would need to complete a standard data enclave training course that would accomplish several goals. The course should provide all users with a course on the responsibilities inherent in accessing sensitive microdata. It should also train users on the logistics involved in utilizing the data enclave. Finally, it would train them on the software tools and other resources provided by the data enclave.

The training course should be designed to support the needs of experienced users, with an abbreviated overview, as well as less experienced users who would need to take more time to go through some or all of the course material.

Clearly, the standard data enclave training course would be augmented by agency-specific and data set-specific training modules. These modules would address issues such as the mechanics of working with a specific data set, as well as confidentiality and access right issues particular to that data set.

Depending on the specifics of the course content and the agencies' requirements, a variety of vehicles could be used to deliver these training sessions. Onsite training could be provided at the enclave. Web-based training could be offered through live Web

seminars or self-paced training software. Finally, the data enclave could take advantage of instances when large numbers of users would be in a single location to deliver in-person training workshop, such as academic conferences or the National Bureau of Economic Research's Summer Institute.

5.5. Dissemination to Academic Community

To maximize the uptake of the data enclave's resources by its intended audiences, consideration would need to be given to marketing to the external researchers. The data enclave and its customers would need to jointly market the value that the data enclave offers its intended researcher audience. In other words, the data enclave and its customers can not simply cite the Field of Dreams principle that "if you build it, they will come." Instead, the data enclave would need to take an entrepreneurial approach to publicizing its existence, resources and unique value to researchers.

The data enclave's Web site would serve as the primary gateway for both potential and current users. The enclave would develop agency-specific pages on its Web site, so that when a user follows a link from an agency's Web site to the data enclave's Web site, it would emphasize that the data enclave is operating as an authorized extension of the agency. The data enclave's Web site would contain content that explains the role of the data enclave, details for potential users the process of gaining access to the archive, and present case studies that illustrate the proven value of the enclave to external researchers.

It is expected that the data enclave would maintain a prominent presence at national meetings and conferences for organizations such as the American Statistical Association, the Population Association of America and the Joint Statistical Meetings. Participation in these conferences could include conference presentations, conference posters and exhibit booths. Ideally, these outreach efforts would be delivered by a collaborative team of data enclave staff, agency staff and external researchers who have utilized the data enclave.

The data enclave could also engage in outreach through the existing communities of data librarians, data archivists and confidentiality researchers, such as ICPSR's Official Representatives (OR) network, the membership of International Association for Social Science Information Service and Technology (IASSIST), and the FCSM's Committee for Data Access and Confidentiality (CDAC).

Finally, the data enclave, its federal customers and possibly highly-motivated external researchers can collaborate to further reach the academic community through low-cost, online distribution channels such as listservs and Web-based discussion forums, blogs, and RSS feeds. These channels could be used both to publicize the launch of the data enclave, and to continually market new resources as they become available, such as new data sets.

3.6. Agency-Specific Data Protection Requirements

Depending on the level of data protection required by the customer agencies, a variety of technologies and techniques can be used to ensure researcher compliance with their data agreements.

For example, the enclave should maintain a complete audit log of all data queries for future review by the data enclave staff. The enclave should control how results sets are delivered to users. Results sets could be delivered to researchers in real time, or only released after automated review by advanced data analysis applications or human review by data enclave staff. If necessary, access to specific features of the data analysis tools such as specific SAS commands could be restricted.

In addition to the technology protections, the data enclave would collaborate with its partner agencies to ensure that it enforces the legal protocols for any specific agency and/or data set. For example, the data enclave could manage the process of external researchers signing a data license before being provided with access to the data.

Finally, as discussed earlier in this proposal, agency-specific and data set-specific training modules could be created that would ensure that researchers fully understand the obligations they are assuming by accessing the restricted data.

3.7. Consortium Interaction

Ideally, the data enclave should provide a shared service to a variety of federal departments and agencies. A major function of the enclave would be to bring together these federal customers, distill their goals and requirements, and then create synergies between them. The goal of this effort would be to create a service that allows the sum of these agencies' investments to be greater than their individual contributions.

To accomplish this goal, the data enclave would need to create a consortium that brings together its customers to communicate and collaborate in a highly-engaged partnership. By taking an active role in this forum, the federal agencies and departments would be able to take part in the ongoing assessment of the enclave, and guide its future efforts.

Each participating statistical agency or department would have a representative member of this working group. In addition, a few additional working group members might be selected from the ranks of the member agencies to bring specific experience in areas such as data processing and dissemination software, computing security, statistical protection methods and end user support for external researchers.

In addition to soliciting feedback from the working group, the data enclave would also need to develop mechanisms for soliciting feedback from the external researchers. A number of cost-effective options exist for collecting this feedback, such as administering a customer satisfaction form, conducting formal and informal customer interviews, conducting user research and usability testing that evaluates the current and potential future offerings of the data enclave. This user feedback data would then be incorporated into both the short-term operations and long-term planning of the data enclave.

4. Summary

This paper outlines a data enclave approach that attempts to combine many of the salient features outlined at the recent NSF SBE/CISE workshop on cyberinfrastructure to develop a data enclave for federal statistical agencies. The approach taken here begins to build the appropriate physical infrastructure that addresses the key cyberinfrastructure challenges identified above – notably the development of:

1. Usable interfaces to support a broad variety of personal privacy policies, and to express trust decisions;
2. Methods for storing, sharing, and aggregating data that are meaningful and yet respectful of integrity and privacy; and
3. Effective forensic tools for detecting and tracking malevolent activity, yet respectful of privacy.